TezProposal

Tez

Abstract

Tez is an effort to develop a generic application framework which can be used to process arbitrarily complex data-processing tasks and also a re-usable set of data-processing primitives which can be used by other projects.

Proposal

Tez is a proposal to develop a generic application which can be used to process complex data-processing task DAGs and runs natively on Apache Hadoop YARN. YARN is a generic resource-management system on which currently applications like MapReduce already exist. MapReduce is a specific, and constrained, DAG - which is not optimal for several frameworks like Apache Hive and Apache Pig. Furthermore, we propose to develop a re-usable set of libraries of data-processing primitives such as sorting, merging, data-shuffling, intermediate data management etc. which are necessary for Tez which we envision can be used directly by other projects.

Background

Apache Hadoop MapReduce has emerged as the assembly-language on which other frameworks like Apache Pig and Apache Hive have been built. However, it has been well accepted that MapReduce produces very constrained task DAGs for each job which results in Apache Pig and Apache Hive requiring multiple MapReduce jobs for several queries. By providing a more expressive DAG of tasks for a job, Tez attempts to provide significantly enhanced data-processing capabilities for projects like Apache Pig, Apache Hive, Cascading etc.

Rationale

There is an important gap that Tez fulfills in the Apache Hadoop ecosystem of allowing for more expressive task DAGs for data-processing applications such as Apache Pig, Apache Hive, Cascading etc.

With emergence of Apache Hadoop YARN, there is a strong need for a common DAG application which can then be shared by Apache Pig, Apache Hive, Cascading etc.

Initial Goals

The initial goals for this project are to specify the detailed requirements and architecture, and then develop the initial implementation including the DAG ApplicationMaster to run natively inside Apache Hadoop YARN.

Current Status

Significant work has been completed to identify the initial requirements and define the overall system architecture. There is a patch available in the internal Hortonworks git repository which can act as the initial seed.

Meritocracy

We plan to invest in supporting a meritocracy. We will discuss the requirements in an open forum. Several companies have already expressed interest in this project, and we intend to invite additional developers to participate. We will encourage and monitor community participation so that privileges can be extended to those that contribute.

Community

The need for a generic DAG application for data processing in the open source is tremendous, so there is a potential for a very large community. We believe that Tez's extensible architecture will further encourage community participation. Also, related Apache projects (eg, Pig, Hive) have very large and active communities, and we expect that over time Tez will also attract a large community.

Core Developers

The developers on the initial committers list include people very experienced in the Apache Hadoop ecosystem:

- Alan Gates <gates at apache dot org>
- Arun C Murthy <acmurthy at apache dot org>
- Ashutosh Chauhan <hashutosh at apache dot org>
- Bikas Saha <bikas at apache dot org>
- Chris Douglas <cdouglas at apache dot org>
- Daryn Sharp <daryn at apache dot org>
- Devaraj Das <ddas at apache dot org>
- Gopal Vijayaraghavan <gopal at hortonworks dot com>
- Gunther Hagleitner <ghagleitner at hortonworks dot com>

- Hitesh Shah <hitesh at apache dot org>
- Jason Lowe <jlowe at apache dot org>
- Jean Xu <jeanxu at facebook dot com>
- Jitendra Pandey <jitendra at apache dot org>
- Julien Le Dem <julien at apache dot org>
- Kevin Wilfong <kevinwilfong at apache dot org>
- Mike Liddell <mike dot lidell at microsoft dot com>
- Namit Jain <namit at apache dot org>
- Nathan Roberts <nroberts at yahoo dash inc dot com>
- Owen O'Malley <omalley at apache dot org>
- Robert Evans <bobby at apache dot org>
- Siddharth Seth <sseth at apache dot org>
- Tom White <tomwhite at apache dot org>
- Thomas Graves <tgraves at apache dot org>
- Vikram Dixit <vikram at apache dot org>
- Vinod Kumar Vavilapalli <vinodkv at apache dot org>
- William Graham <billgraham at apache dot org>

We realize that though we have significant employer diversity already, additional diversity is always better, and we will work aggressively to recruit developers from additional companies.

Alignment

The initial committers strongly believe that a standard task DAG application on Apache Hadoop YARN will gain broader adoption as an open source, community driven project, where the community can contribute not only to the core components, but also to a growing collection of applications which will be based on top of Tez. Our hope is that the Apache Hive, Apache Pig, Cascading and other communities will find tremendous value in Tez and will adopt it en masse.

Known Risks

Orphaned Products

The contributors are leading users and vendors in the Apache Hadoop ecosystem, with significant open source experience, so the risk of being orphaned is relatively low. The project could be at risk if vendors decided to change their strategies in the market. In such an event, the current committers plan to continue working on the project on their own time, though the progress will likely be slower. We plan to mitigate this risk by recruiting additional committers.

Inexperience with Open Source

The initial committers include veteran Apache members (Committers, PMC members and Apache Members) and other developers who have varying degrees of experience with open source projects. All have been involved with source code that has been released under an open source license, and several also have experience developing code with an open source development process.

Homogenous Developers

The initial committers are employed by a number of companies, including Cloudera, Facebook, Hortonworks, Microsoft, Twitter and Yahoo. We are committed to recruiting additional committers from other companies based on their contributions to the project even though we do have significant diversity already.

Reliance on Salaried Developers

It is expected that Tez development will occur on both salaried time and on volunteer time, after hours. The majority of initial committers are paid by their employer to contribute to this project. However, they are all passionate about the project, and we are confident that the project will continue even if no salaried developers contribute to the project. We are committed to recruiting additional committers including non-salaried developers.

Relationships with Other Apache Products

As mentioned in the Alignment section, Tez is closely integrated with Hadoop, Hive and Pig in a numerous ways. We look forward to collaborating with those communities, as well as other Apache communities.

An Excessive Fascination with the Apache Brand

Tez solves a real need for generic task DAG management in the Apache Hadoop ecosystem, something which has been addressed in a very ad hoc manner so far by multiple Apache projects. Our rationale for developing Tez as an Apache project is detailed in the Rationale section. We believe that the Apache brand and community process will help us attract more contributors to this project, and help establish ubiquitous APIs.

Documentation

http://wiki.apache.org/incubator/TezProposal

Initial Source

Available as a patch.

Cryptography

Tez will eventually support encryption on the wire. This is not one of the initial goals, and we do not expect Tez to be a controlled export item due to the use of encryption.

Required Resources

Mailing List

- tez-private
- tez-dev
- tez-user

Subversion Directory

Git is the preferred source control system: git://git.apache.org/tez

Issue Tracking

JIRA Tez (TEZ)

Initial Committers

- Alan Gates <gates at apache dot org>
- Arun C Murthy <acmurthy at apache dot org>
- Ashutosh Chauhan <hashutosh at apache dot org>
- Bikas Saha <bikas at apache dot org>
- Chris Douglas <cdouglas at apache dot org>
- Daryn Sharp <daryn at apache dot org>
- Devaraj Das <ddas at apache dot org>
- Gopal Vijayaraghavan <gopal at hortonworks dot com>
- Gunther Hagleitner <ghagleitner at hortonworks dot com>
- Hitesh Shah <hitesh at apache dot org>
- Jason Lowe <jlowe at apache dot org>
- Jean Xu <jeanxu at facebook dot com>
- Jitendra Pandey <jitendra at apache dot org>
- Julien Le Dem <julien at apache dot org>
- Kevin Wilfong <kevinwilfong at apache dot org>
- Mike Liddell <mike dot lidell at microsoft dot com>
- Namit Jain <namit at apache dot org>
- Nathan Roberts <nroberts at yahoo dash inc dot com>
- Owen O'Malley <omalley at apache dot org>
- Robert Evans <bobby at apache dot org>
- Siddharth Seth <sseth at apache dot org>
- Tom White <tomwhite at apache dot org>
- Thomas Graves <tgraves at apache dot org>
- Vikram Dixit <vikram at apache dot org>
- Vinod Kumar Vavilapalli <vinodkv at apache dot org>
- William Graham <billgraham at apache dot org>

Affiliations

The initial committers are employees of Cloudera, Facebook, Hortonworks, Microsoft, Twitter and Yahoo Inc.

- Alan Gates - Hortonworks
- Arun C Murthy - Hortonworks
- Ashutosh Chauhan - Hortonworks
- Bikas Saha - Hortonworks
- Chris Douglas - Microsoft
- Daryn Sharp - Yahoo
- Devaraj Das - Hortonworks
- Gopal Vijayaraghavan - Hortonworks
- Gunther Hagleitner - Hortonworks
- Hitesh Shah - Hortonworks
- Jason Lowe - Yahoo
- Jean Xu - Facebook
- Jitendra Pandey - Hortonworks
- Julien Le Dem - Twitter
- Kevin Wilfong - Facebook

- Mike Liddell - Microsoft
- Namit Jain - Facebook
- Nathan Roberts - Yahoo
- Owen O'Malley - Hortonworks
- Robert Evans - Yahoo
- Siddharth Seth - Hortonworks
- Tom White - Cloudera
- Thomas Graves - Yahoo
- Vikram Dixit - Hortonworks
- Vinod Kumar Vavilapalli - Hortonworks
- William Graham - Twitter

The nominated mentors are employees of Hortonworks, LinkedIn, NASA JPL and Microsoft.

- Alan Gates - Hortonworks
- Arun C Murthy - Hortonworks
- Chris Douglas - Microsoft
- Chris Mattman - NASA JPL
- Jakob Homan - LinkedIn
- Owen O'Malley - Hortonworks

Sponsors

Champion

Arun C Murthy <acmurthy at apache dot org>

Nominated Mentors

- Alan Gates <gates at apache dot org> ,Äì Architect at Hortonworks. Committer for Pig.
- Arun C Murthy <acmurthy at apache dot org> ,Äì Architect at Hortonworks. Committer for Hadoop.
- Chris Douglas <cdouglas at apache dot org> - Sr. Research Engineer at Microsoft. Committer for Hadoop.
- Chris Mattman <mattmann at apache dot org> - Sr. Computer Scientist, NASA JPL. Committer for Nutch, OODT and Tika.
- Jakob Homan <jghoman at apache dot org> ,Äì Sr. Software Engineer, LinkedIn. Committer for Hadoop, Kafka, Giraph.
- Owen O'Malley <omalley at apache dot org> ,Äì Architect at Hortonworks. Committer for Hadoop, Ambari.

Sponsoring Entity

Incubator